THE BROOKINGS INSTITUTION
WEBINAR

OPERATIONALIZING RESPONSIBLE AI

Washington, D.C.
Tuesday, April 5, 2022

PARTICIPANTS:

DARRELL WEST, Moderator
Vice President and Director
Governance Studies
The Brookings Institution

ELIZABETH ANNE WATKINS
Postdoctoral Research Fellow
Princeton Center for Information Technology Policy
and Human-Computer Interaction
Princeton University

QIAN YANG
Assistant Professor
Computing and Information Science
Cornell University

MEGAN YOUNG
DLI Postdoctoral Fellow
Digital Life Initiative
Cornell Tech

* * * * *

P R O C E E D I N G S

MR. WEST: Good morning. I'm Darrell West, the Vice President of the Governance Studies at the Brookings Institution. And I'm pleased to welcome you to our webinar on operationalizing responsible AI.

So there's wide spread agreement among ethicists that responsible AI principles require fairness, transparency, safety, and explainability. But it is not always clear how to operationalize those broad principles or how to handle situations when conflicts arise between them. Moving from the abstract to the concrete when developing algorithms often presents challenges as a focus on one goal can come at the detriment of alternative objectives.

We have a new Brookings report out this week entitled six steps to responsible AI in the federal government and you can find that paper online at brookings.edu with just a quick summary of our six recommendations. There's one having concrete codes of conduct.

Secondly, having appropriate operational tools for promoting major ethical principles and fighting bias. Three developing clear evaluation benchmarks and metrics for relying on technical standards to help with common problems. Five, experimenting the pilot projects in organizational sandboxes. And the finally, having a mix of technical and nontechnical skills in the workforce so that people actually can operationalize them.

To help us understand these and other issues related to technology innovation, we're delighted to have three distinguished experts with us. Meg Young is a Postdoctoral Fellow at the Digital Life Initiative at Cornell Tech, which is based in New York City. Qian Yang is an Assistant Professor in Computing and Information Science at Cornell University. And Elizabeth Watkins is a Postdoctoral Research Fellow at the Princeton

Center for Information Technology Policy and Human-Computer Interaction at Princeton University.

Now, if you had questions for our panelists, you can email them to us at events@brookings.edu, that's events@brookings.edu or you can Twitter at @BrookingsGov using the #ResponsibleAI. So that's Tweeting at @BrookingsGov using the #ResponsibleAI.

So what we're going to do today is discuss ways to operationalize responsible AI and move towards more concrete standards. We're going to look at how to design appropriate algorithms and build technical capacity in the workplace.

So I'd like to start with Meg. You argue that it is important that there be multiple stakeholders at the table for AI decisions. And we know there are lots of issues that these algorithms are raising. So who should be at the table? And how should we engage the public on these important issues?

MS. YOUNG: Thank you, Darrell. I'd like to say that I'm delighted to be here.

What I've noticed in working with municipal governments is that often it's the same usual suspects that we bring to the table. Those who are very engaged in technology policy and they therefore are going to present, but there's a concern that I've heard that the general public doesn't have the technical expertise to inform these important conversations.

However, in the academic community we recognize that people who have lived experiential, people who are on the other side of the AI systems are exponential experts and they have a lot to offer. We know that talking to them can include users of the system like field operators, people who collect the data that the system is using or those who are at the other end of the chat bot or system decisions.

And the system isn't just the hardware or the software. It's also these

policies and norms that are shaped by users in their everyday work practices. So if you're not talking to users, there's going to be unexpected performance issues and you won't be aware of their needs and harms. So how do we do that?

I mean it's all well and good to say, here's the public but what does it actually look like. We need to start where we are. So even though this is a very technically complex topic at first blush, when you talk to people with limited experience, they can share their questions about a given technology and their concerns.

I was recently at a library hosted event at NYU by Eric Corbett that asked the public, what would you want to know about the system? And that can help take what is an enormous amount of information. Agencies are having trouble prioritizing what to share out to the public and compress it to just those high priority items that the public says that they want and need to know.

This can also help inform reporting processes that we've seen are too lengthy to work well. So although, I have other points about this closing the loop between program managers and the public can be a good in itself where these relationships don't exist right now. It's important to listen to them and it can attune us to the sensitivities of the people who are most effect by the system.

MR. WEST: Those are all great points and I love that concept about experiential expertise meaning we often think it's the technical people who have the expertise. And so, we should delegate all of the decisions to them. And I think you're making an important point. People do have personal experiences. They have lived experiences. Their perspective actually is important as we start to operationalize a responsible AI.

So, Quian, in many areas we still are in the early stage of AI adoption. And this is certainly the case in the federal government and many federal agencies. What are

the risks at this point in time? And how can we mitigate them?

MS. YANG: That's a great question. Thanks, Darrell. I think in short, I think the biggest challenge in adopting AI is to keep AI unremarkable. Keep AI support to human desires, human needs and organizational purposes.

I often make an analogy that AI is going to be the electricity of our time. It is important for us to know that electricity is a clean technology, but with the right interface design, infrastructure design and policy design we can keep technologies useful and safe to use for people and for societies. And that's what the challenges we're dealing with in a lot of the AI application areas.

For them to work well in healthcare where you see doctors often rejecting the idea of AI because of the many concerns about ethics and biases and all the other things. But you also see that there's sometimes -- sometimes, you see people accept AI in their practices too soon. You see once there's an official prediction that you're commended by the government agencies, we see doctors do not start their decision making without the AI on the table. And there's a real center around their work practices around AI. Then that is the tricky balance we're dealing with. How do we make sure AI plays the role, but not too significant a role such that its problems will permeate in our actual practice, healthcare practices?

And that is the challenge, I think, we're still learning to deal with and try to make sure we engage with the public. We engage with the users, the doctors in the healthcare case. That engage any stakeholders that are part of the process, but also have a constructive and effective conversation between AI experts and these people in designing the policies and interfaces of the systems around this kind of AI.

MR. WEST: Yeah. Having the right interfaces, I think is crucial. And I know each of the three of you, you do talk about human/computer interactions and how we

get that right. And so, that is certainly important in going forth.

Elizabeth, I know you put a lot of emphasis on accountability in AI algorithms. What does this mean? And how should that factor into AI development?

MS. WATKINS: This is a great question, Darrell. Seconding the thanks of other folks who have said. I'm very privileged to be on this panel with these amazing thinkers. Seconding everything that everyone has said about the public and engaging the public.

And as we think about how we can make not just algorithmic systems accountable and how to make the companies and firms that produce accountable. But we need to think very clearly about the structures that we build in terms of the power that we grant to the public and the relationships that we build between the public audiences or the users who are the ones actually consequenced and harmed by these systems. And the ones who like makes it have experiential expertise in how these systems are being deployed into sociotechnical context in communities and work and health.

But also, how we empower these folks to have mechanisms for redress when things go wrong or when things don't perform as anticipated or when dual uses arise or when malicious uses arise. And there's a tool that's really powerful towards that end and that's a conversation that a lot of people have been engaging in is around algorithmic impact assessment. And there's a lot of tools for accountability that people have been talking about in terms of, for example, like audits.

Audits are really powerful. And a really great way to check to see if the system is doing the things that it's supposed to do. But when it comes to giving the public a mechanism to say, hey, this system isn't doing the thing that it was supposed to. Or, hey, we can see -- and for example, this privacy impact assessment that our data is being handled in a way that is unsafe or that might present novel harms. Then we need to

empower the public to both access and be able to scrutinize documentation about how these systems are built, how they're designed and how they're governed.

And so, that's going to take ongoing building of -- like you say, Darrell, in the Brookings' report -- building of workforce capacity and empowering folks who are in the government to be able to build mechanisms for algorithmic impact assessment that can encompass things like social science expertise so that social scientists can go and talk to the users and analyze the kinds of harms that are happening in communities that might not be anticipated by a tool like an audit.

MR. WEST: And I do like the idea of impact assessment and audits. And we talked about them in our work as well because sometimes in algorithms there are unappreciated consequences. I mean designers may have perfectly good motives. They develop an algorithm but then it ends up doing things differently than what they intended.

So keeping track of that, monitoring the impact, doing audits, those are all ways that federal agencies can start to move into the right direction.

So Meg, now that we're on this topic of federal agencies. How should federal agencies develop guidelines for responsible AI? Do they need codes of conduct? Are there operational guardrails they can put in place that would increase the odds of the AI acting in a responsible manner?

MS. YOUNG: Thanks, Darrell. Absolutely. I have noticed that one of the strongest first steps that an agency can take is either to have a policy or a public commitment and a strategy to a set of digital rights principles that can commit them to taking action on a responsible AI.

But that's only a first step. And a lot of the work that's been coming out in the last few years has been trying to bridge that gap from making the commitment to taking action. And I think that while often this is considered to be a technical problem and one that

I'm sure that we'll talk about. I could go more into depth about.

It's really not only a technical problem. And even agencies that are still trying to bolster their technical expertise can still take action. Now, your report does an awesome job of highlighting the organizational and policy factors the agencies could adopt, but what exactly are those?

It's a matter of defining what are the policies that govern a system in use? What training do operators get about, for example, the error rates or the system accuracy? And most important of all are the things that an agency does to communicate with the public about the system.

There's a research, nonprofit in D.C., Upturn, that found that you can have a very meaningful public dialogue about a system with just a few things that most of the agencies listening could publish in the next few weeks.

It is the existence of a system. Its purpose. The policies that govern its use and the system inputs and outputs. These simple things are a good starting point for a dialogue. And it shows how much agencies can do.

And just a final thought there. In our work, we found that advocates want something as simple as the name and contact information of somebody that they can take their questions, concerns or recourse to for systems that are not working correctly.

MR. WEST: Qian, I'd like to get your thoughts on this very same question. How should federal agencies develop guidelines? What kind of operational measures should they consider in order to increase the odds of responsible AI?

MS. YANG: Thanks, Darrell. I want to just echo everything everyone just said. I think that these are really good points.

I also want to add that I think AI is really a nebulous term covering a really wide range of technologies that are actually fundamentally really different. And we need the

kind of technical expertise to understand at what point we need to separate. We need to differentiate these technologies when we regulate them and when we design them.

One example is for example the language. When we talk about language technologies, we're seeing foundation models becoming a new thing and it really changes how systems are built. When foundational language model contains billions of perimeters that can function across many, many domain areas from writing, to conversation, to writing programming codes to do math. That really requires -- that kind of foundational technology reminds me a lot of like genetic engineering or CD printing. This kind of foundational technology that doesn't produce one thing, but what it produces depends on what kind of data that you put in. And that I think -- that kind of technology I think requires a different way of thinking when we regulate them.

Another aspect I want to just point out is I think when we regulate AI, there's the algorithmic fairness, but there's also data fairness. And data representation in the fairness are about human beings. They're about people who generate those data and the power of AI lies in the fact that it can learn life from its users. It's the AI's greatest power, but it also one of its biggest risks.

And I don't think we quite know how to regulate and how to foresee all the possible ways an AI can go wrong when it learns from its users constantly. And I think that's still a challenge. But I sort of dividing and conquering this giant thing called AI, I think we can make headways in regulation and in responsible AI as well.

MR. WEST: It is important to emphasize the data aspect of the algorithm because often times the data that go into the algorithms are either incomplete or unrepresentative. Either of which can create major problems in terms of the actual outputs and the actual decisions that get made. So we certainly need to pay attention to that.

Elizabeth, how do you think federal agencies should develop guidelines?

What can they do to operationalize responsible AI?

MS. WATKINS: Yeah, thank you. So there's two points I want to make. So one is by returning to the template of impact assessments. Something that is really exciting about impact assessments is in their history and why they were developed in the first place.

And impact assessments were developed in order to give a tool to communities for power and redress. Folks who didn't have other kinds of administrative power. Folks who were getting consequenced in particular by developments of large-scale engineering projects that might have environmental impacts on their water or on their air.

And so, the power of these impacts was additional in the fact that they were tied to robust judicial and legislative backstops. That people could see the impact assessment and they could use that and take it to courts and say, hey, this environmental impact assessment says there has to be a specified threshold of pollution and what's actually being created or what I'm actually experiencing exceeds that threshold. So this needs to be redressed. This needs to be redesigned.

And this combination of public's access to documentation, of public's ability to use that documentation to see redress through the courts and the presence of scientific expertise to identify what potential harms are. Those all come together into a web of relationships. And it's that web of relationships that turns into true accountability.

So as the government is thinking about guidelines. Pursuing the kinds of guidelines that can foster and facilitate those kinds of relationships is so important. And also, to build on the conversation about the distinction between AI and some algorithms. And I see that we have a question about preprogrammed AI algorithms that might be something as simple as a decision tree or a regression model in a model that might be based on something like deep learning.

Where an explanation perhaps can't be seen. Or that the conclusions that

are being made or the predictions that are being made by a model might be less than transparent. This also needs to be heavily considered in federal guidelines around which types of systems are being used? In particular, in higher risk scenarios like work and safety and infrastructure?

And I had a conversation yesterday with a computer scientist at Princeton named Angela Wang. And she pointed out that if there's a model that we don't understand then we shouldn't have it in a high-risk scenario because a lot of folks criticize sometimes the ability of decision making from algorithmic models when it's compared to human decision making.

Well, humans make mistakes too. So what does it matter if the AI makes mistakes? But with human mistakes, we have procedures for how to inquire into their decision making and for how to correct or address mistakes that are being made or errors that are being made. But so far with AI models, we don't. And that's where that distinction is. It's what comes after the accident happens? It's what comes after the harm happens?

MR. WEST: So, Meg, one key problem in the federal government is an inadequately trained workforce. What can we do to improve worker training in regard to AI deployment? And how can we provide people with the skills needed for responsible AI?

MS. YOUNG: Thanks, Darrell. Absolutely a boosting responsible AI skills and experience is an essential part of this equation.

My first suggestion is that digital rights belong in your org chart. I have noticed organizations that operationalizing effectively are creating a central hub that is responsible for these conversations and skilling up. And creating tools and templates that program managers can use who are closer to the operational mission of a specific deployment. Together they are better at thinking through the ethical implications of AI at every step.

And when you have a central hub, it's also easier to do learning throughout the organization or to coordinate across agencies. You can imagine picking up the phone and talking to another agency that has used the same vendors as you before.

I also think that external experts are really an important part of this equation. I know that there are programs in the federal government that allow experts to do a tour. And I think that in addition to these existing programs, I would love to have agencies explore the idea of a blanket research agreement to bring in academic researchers under the hood in a more frictionless fashion who could advise, collaborate, evaluate pilots.

Often times legal barriers are a factor that slows them down, but looking at examples like the city of Austin that has partnered with the University of Texas, Austin. They're able to collaborate on a number of digital rights initiatives under one blanket, service agreement. And I think that this could be a great model for agencies to access that external technical expertise while learning populates through the organization.

MR. WEST: Yeah, that is a great point because there's tremendous expertise out there outside of the federal government. And certainly, in academia. The three of are great representatives of that. And so, if there are ways to facilitate the partnership between academia and the federal government that certainly could be a way to get trained people into these processes.

Qian, the same question for you. What are your thoughts on federal workforce? How we improve the training? What are the types of skills that people need? And how can we go about providing them with those skills?

MS. YANG: Yeah. Great question. I think Meg had mentioned quite some of those are really valuable. I would ask that besides the technical AI expertise and the human understanding with the experience kind of expertise.

We really need workforces that know how to collaborate across these

disciplinary lines. We did quite some research into AI product design, development teams in the industry across many, many companies. We see a struggle between -- like to collaborate across these expertises. Data scientists or AI experts say, tell me what is the fairness and evaluation metrics so I can optimize my model towards that, quote, unquote, fair or whichever adjective you use for describing responsible AI.

I will optimize my models for it. And then the human experts say, well, it really comes back -- it's really nuanced. It depends on how people feel about this. And there are many ways they could do this. So these kind of conversations often hit a wall even in the early stages.

And we're seeing those kinds of tension play out in many organizations. And I think that is a challenge we're seeing. And one of the greatest privileges in working at a university is that we are seeing more and more students coming out who are trained in both aspects. Who can both speak the language of quality type of research, human understanding and algorithms and deep learning? And that I think is another critical aspect to it.

And I think policy is another aspect. I mean, Meg and Elizabeth can speak much more to that. But I think policy, the boundary between design policy is becoming quite blurred. But I think that is another area we also see needs a lot of interdisciplinary expertise.

MR. WEST: Yeah. And I think that's exactly right that we need the technical and the nontechnical capabilities. I mean, the human element is crucial in the design and just the sociological consequences of algorithms and how they're making decisions. So certainly, when we talk with people in federal agencies, we emphasize a mix of skills that's going to be necessary and the importance of kind of working across subject area of values so that you get the right expertise in them.

Elizabeth, I'd love to get your thoughts on these questions of workforce

development. How it can improve the workforce? How we can give them the skills needed for responsible AI?

MS. WATKINS: Sure. This is a great question and I'm really glad that thank you, Qian, for that lead in on the line between design and policy getting blurred.

I think this is really important not just for the skills that are being brought into and developed within the public sector, but also doing what we can to go the other direction. And to take the expertise within the public sector about how tools are used? How they are working with and serving their publics? And getting the expertise back to the designers and to the vendors who are the ones who are building and designing these systems.

And I just read a really compelling paper by Damon Sucheta who is one of the authors. And he wrote about risk prediction systems for child welfare caseworkers. And there was a really compelling example that he and his coauthor wrote about where there is an algorithmic system that determines how much a foster family should be compensated for their care for a child that's awarded to the state.

And the compensation is based, in part, on how much the child, for example, is experiencing anxiety or acting out that if a child needs more care then the family will get compensated more. But the probably when foster families, when they do the work and they put in the care and safety that these children need. And the children start to get better and their behavior starts to improve that means that the families get compensated less over time. So they're actually getting penalized for doing a wonderful job with these kids.

And that becomes a problem with the caseworkers who are working with this system. And the caseworkers have to come up with different kinds of data input methods to make sure that these families can still afford to care for these children at all because there were cases documented where families had to actually stop caring for these

children because they just didn't have the resources even though they had put in the work to get these kids better.

And so, getting more expertise about how these systems were used and how they impact the public and getting the expertise back to the builders would be a fantastic step towards all kinds of workforce capacity building.

MR. WEST: So I have one more question for our panelist and then we will move to audience questions. We're already getting a number of really interesting questions from our viewers.

So my question is on the best ways to move forward with AI? Because we know there are problems of fairness, bias, transparency, human safety and explainability. How do we operationalize these principles? What happens if there are conflicts among these important principles? And, Meg, we can start with you on that.

MS. YOUNG: Thanks for the question. I think that the most important thing to operationalizing these ethical principles that we haven't spoken about yet is technology procurement. Procurement processes are an amazing lever that the government has at its disposal to try to bake these values into the technologies.

And it can be challenging because when a vendor is responsible for the system, the agency might have limited visibility or reach into the vendor's process. But there have been a lot of smart people thinking about this and resources that are getting developed for arming agencies with the right questions to ask.

I would refer you to Rashida Richardson at Rutgers or Mona Sloane (phonetic) at NYU. And their recent primers on how the procurement process is a critical moment where we can begin to bake in questions about fairness which as you noted in your report, Darrell, are not straightforward at all.

There are 20 different definitions that could apply for fairness, it is another

critical moment to engage with your stakeholders, your end users, your people who are affected. And to ask them which one makes the most sense for this context? Once you have the answers to that this is information that needs to be brought back to the vendor. You can define at that negotiation step additional features, evaluation needs or documentation that you can compel the vendor to provide for your agency.

And longer term as these frameworks that were initiating become more stabilized, this can be baked into the standard procurement process the same way that cyber security checks or legal compliance checks are.

And once you have some assessment that you're making, this is the kind of information that should be shared back out to the public so the public knows how this thinking about digital rights and responsible AI inform the design of the system that you're now using.

MR. WEST: So, Qian, your thoughts on how to operationalize these principles? And then what happens if there are conflicts among the principles?

MS. YANG: Great question. I think, I'm not sure of a pessimistic view which is I think there will always be conflicts. And because there were so many different AIs in so many different like life causes. I think it will be difficult to intermarry them all or anticipate them all before you deploy it.

And I think one thing -- there are two things that I think we can do immediately. One is I think to recognize that not too (inaudible) and high PS AI is the most usable for us. I always recommend AI designers to start with the simplest thing. If you want to create a computer vision algorithm that does, you know, diagnosis cancer on images. Start with a Rubik system and test how the doctors reacted. How the patients will react. Were there possible societal consequences there are.

And before you add onto the explainability challenges and computer vision

and deep think learning. Those additional challenges that even the complexity around the binary rule-based classifier is substantial. And I think always starts with the small and simple one.

And the second thing I wanted to say is leveraging existing tools in the past. I think from a policy and regulation perspective, it is also we are seeing AI technologies becoming more homogeneous in a way. There's a fixed set of models and algorithms that many, many different kind of AIs read out of.

And I think we can by regulating and making sure these tools in this foundation models are reliable, are fair, responsible. I think we can make faster progress in that direction.

MR. WEST: Elizabeth, your thoughts on operationalizing these principles and what happens if there are conflicts among them? And then after this we're going take questions from the audience. You can email us events@brookings.edu.

MS. WATKINS: This is such a tough question. I agree with the other panelists. There are always going to be conflicts in particular with technologies that touch so many lives. And that can have such crucial critical life changing impacts on people. Like being arrested or being granted a housing loan or being granted credit or getting hired for a job.

And so, I want to second what Quian said that the starting question when we're developing a product really shouldn't be how do we build it, but should we build it? And the AI isn't always the best tool for the job. And there are lots of other tools that we have to rely on that are already existent within guardrails and within regulatory processes.

And in the recent work on impact assessments that I've been lucky to contribute to with data and society team on AI on the ground. We've recommended that any effort at either building or governing systems that are going to touch so many people's lives

engage in public consultation with all kinds of different people of varying kinds of expertise. And if the question ultimately comes to a point of conflict but that the harms that might be experienced are so high risk or so high stakes then the idea that this tool simply not get built or not be deployed, needs to be on the table.

MR. WEST: Okay. Thank you. So now, we're going to take some questions from the audience. And again, you can email us your questions at events@brookings.edu or Tweet at @BrookingsGov using the #ResponsibleAI. And we will get to as many of your questions as possible. So here's the first question. And any of you who want to jump in are welcomed to do so.

AI always has a degree of uncertainty associated with it. And this person basically says, therefore AI should not be making some decisions but should only be a source of information for human decision making. And if I could just add a quick comment to that. My sense of what the person is asking is perhaps we shouldn't have AI making autonomous decisions, but only being an input into human-based decision making. I'm just curious. Any of you have any thoughts or comments on that?

MS. YANG: I think that is a very good point and I will add two more complexities to this question.

I think autonomous driving as an example. It is actually much easier to develop a fully autonomous car rather than designing a car that codrives with a human driver. Just imagine you could drive a car with even your best friend who really understands you.

So I don't think -- let us just say, I don't think the distinction between fully -- decisions that should be fully automated versus not are merely a distinction between how the important decisions are. I think it models the task itself and the humans that are involved in it.

And the second point I wanted to make is actually it sort of leads back to what Elizabeth was saying earlier. Like an explainable algorithm shouldn't be used in practice. I also don't think it's always true. We did a lot of work in, for example, computer revision in healthcare. I mean it is totally possible for doctors to make a good decision with the help of an AI in reading medical images. It is really in our experiments we could repeatedly show that even if the model isn't expendable when you design AI carefully, it can nonetheless trigger doctors to think more. It can challenge the doctors to think more carefully about their cases.

The challenge however is in practice how can we make sure the AI constantly changes the doctor without the doctor being annoyed and think I will just not use this if you ask me to rethink my decision every time. Looking at the nuances in terms of the convenience and automation brains versus the risks. It is always easier to think, oh, we will just, you know, have the human as the gatekeeper. But the human beings probably don't want to be the gatekeeper of AI all the time.

So it is a really important and difficult question, but I fear I don't have a conclusive answer to this.

MR. WEST: Meg and/or Elizabeth?

MS. YOUNG: I'd like to pick up on the part of the question that mentioned the degree of uncertainty inherent in AI systems. We know that the accuracy statistics the product is shipped with are often based on tests that are under lab conditions and don't resemble the real-world context that our systems are deployed in.

So there's an increasing move in the academic community to acknowledge that this pressure for systems to be developed in a small environment and then to scale across context is unwise and dangerous. And it's this question inspired to me to think about how we can balance what the appropriate deployment is for a system and to define when a

system is appropriate and not appropriate to be used. What academics have been calling non-scaler systems. Systems that don't scale outside a specific context.

MR. WEST: Thank you. Elizabeth?

MS. WATKINS: Sure. Just to bring up one last point. I'm going to go back to the Sucheta and coauthor paper on the child welfare caseworkers in that there was a really fascinating distinction that they saw between different kinds of caseworkers when they were using the outputs of the system to make decisions about compensation of loss for families.

And they found that the younger caseworkers who were less experienced and had less of their own experience to draw from were far more likely to accept the outputs of the system and to allow the system to hold the liability to say, okay, you know what? That's a decision the system made then I'm going to go with it. However, the older caseworkers who had at least, I think, eight years of experience and therefore many years of experience to draw from, they were more likely to reject the outputs of the system and to say, no. I'm going to use my professional discretion and I'm going to make a different decision.

And so, being cognizant that when we talk about a human in the loop, there is no one objective human. All these humans are different. They're all bringing very different kinds of professional expertise and professional obligations to their interactions is something that I don't see talked about very much. That I think might help us to think about the interaction between the human and the system.

MR. WEST: So we have another question and any of you are free to jump in on this. What are the key policy and technical challenges in operationalizing responsible AI?

MS. YOUNG: I think that a lot of the guidance that I gave earlier is for

agencies that are still early on in their journey as operationalizing responsible AI. But from a policy perspective, more mature organizations might be looking for examples and templates.

And I look to the surveillance ordinances that passed across the country in the past five or six years to see how from a policy perspective you can pass a strong commitment to transparency. In this case, it was for surveillance systems, but you could imagine it for AI. Where an agency has to disclose all the systems that they're using and to report on those systems, do public engagement and to pass a vote, yes or no, on whether to use the system based on that reporting.

I wrote a case study on this with colleagues in 2019. And essentially, we found there to be a few challenges. One, you touch on this in your report, Darrell. Defining what AI is, is an enormous policy challenge. If you have a definition that's too narrow, you might miss technologies that are harmful. And if you have a definition that's too wide, you could end up putting things through this responsible AI process that don't belong like cell phones or, you know, Microsoft Excel. You know, arguably Microsoft Excel required that kind of a process, but you got my point.

A second challenge is that a reporting or documentation process can sometimes get unwieldly. In Seattle, the assessments that they were making of the surveillance impact were sometimes hundreds of pages for a single technology. And that's not very useful for advocates to bite into, engage and give their feedback.

So finding a way to pass policies that enable strong transparency requirements and reporting without creating a mountain of work for agencies and for advocates and the public to dig into I see as one of the key challenges that we can improve on based on our learning from the surveillance ordinance example.

MR. WEST: Elizabeth and/or Qian?

MS. WATKINS: Sure. Thank you. I'll draw on some of the critiques from

the history of privacy assessment. Privacy assessment ideally is a process that designers engage in from the beginning of the design process.

And so, thinking about the risks that their data collection and handling and governance processes might bring to other the company producing that system or to the users who are going to be consequenced by that system. Hopefully, that thinking process then turns into a more human centered and privacy centered design process.

Over time, however, there have been critiques that the privacy assessment has ended up just being a checklist process that happens at the end of the engineering or design cycle where a product is about to be launched or unfortunately it has already been launched and there's just a checklist of like, oh, yes. We protected this. And we saw this correctly. And yes, we did all this.

And so, I think one of the challenges of that Meg has defined in terms of how unwieldy these kinds of guardrails can get for organizations that often have like PPIs and performance reviews and organizational constraints and performance constraints that they have to meet. Making sure that's the guidelines and how they're operationalized can be effective throughout the design process and not just a checklist at the end could make a real difference in terms of the harms that these systems produce.

MR. WEST: Okay. Thank you. We have another question about IP protection and security risks. Obviously, these are particularly important issues in federal agencies.

As federal agencies are developing their algorithms and rolling out AI systems how do we protect intellectual property? And how do we deal with the cyber security risk that particularly in our current environment seem to be growing all the time?

MS. YANG: That is a wonderful question. I can speak to the cyber security aspects and Megan will be more the expert on the privacy and IP front.

As I said, I think more and more AI systems are driven by the same set of algorithms and the same set of foundational technologies. And this trend, I think raises serious concerns and challenges in ensuring security. We are seeing more and more work in ensuring that important models that work in critical decision making are adversarial attack proof.

But then as these systems -- and these are really important, but also other systems that are really rely on user generated data. Your smart watch, your smart phone, the data they track and the models that are built upon those data for example, heart rate or blood monitoring. This kind of algorithms. These are really easy to -- these models are really easy to be attacked for just simply by users turning on and off the sensors in certain ways. And I think that really raises the stake that we're seeing in ensuring model security.

So I think this is another area where human expertise and model expertise, technical aspirations really need to come together and to understand what are the new ways of attacks that might possible because of these sort of AI infused sensing and AI state of monitoring.

MS. YOUNG: I'd like to pick up on the part of the question that talks about trade secret. When I was at the University of Washington, we looked at trade secret barriers to data sharing. And of course, the government has many data access needs for proprietary firms, for example. Transportation agencies wanting to understand how Lyft and Uber moved through public space.

And we found that creating a trusted third party. I was host at the university gave an opportunity for researchers to get under the hood of proprietary systems and to engage public agency questions without the firm's feeling to threatened about their proprietary technology being in a space that was potentially open for public records requests or access by their competitors.

MR. WEST: Okay. So we have a question about the European Union. And in general, on many technology policy issues, the EU has been tougher than at least current American policy especially at the national level.

And so, in regard to AI, the European Union has developed new roles and new guidelines based on risk assessment where they basically try and categorize the AI based on the degree of risk and the number of people who potentially would be affected by the AI. And then they gear the regulation to that level of risk. So the AI could fall in a low risk category and therefore not warrant much regulation, a moderate risk or a higher risk category which then obviously would dictate a higher regulation.

I'm just curious how each of you react to that approach. And also, just the difference between the U.S. generally being a little lighter on AI regulation and the EU being a little tougher on AI.

MS. WATKINS: If it's okay for me to jump in? This is a really great question. There is a big challenge with how AI technologies, algorithmic deep learning, machine learning model-based technologies can use similar core technologies but be deployed in very different domains such as housing or access to financial tools or equal access to employment opportunity or into finance, if I didn't mention that one already.

And these domains tend to be and these sectors tend to be regulated very differently and subject to very different regulatory regimes in particular to the orientation and protection of protected groups and protected categories of people. And so, this presents a big hurdle for the people who are trying to put guardrails around these technologies.

And shifting the frame from sector to risk category, I think is a really advantageous step forward to thinking instead of into which regulatory sector of technology is going to be deployed. Instead thinking how much risk is this going to visit upon vulnerable people and impact communities? I think that's a great step.

MS. YANG: Again, I have a pessimistic view here. I think --

MR. WEST: It's good to have a range of perspectives here. You can say that so.

MS. YOUNG: Indeed, it's a great step forward. And I think there's also important, I think to recognize that risk categories are not the only way. It will not stop all the problems that we're facing.

One example, if you measure a heart rate in the clinic that is HIPAA protected really privacy sensitive data, but if it's your smart watch that sends to your heart rate, it is constant currently categorized as personal/commercial data. And it's accessible for all kinds of machine learning models to mine. And in that sense, I think again like risk kind of in our traditional ways of thinking about risk categories like if you're in the hospital. It's like really sensitive. And if it's not, it's really simplified view of risk and human consequences of AI. And I think that conversation about AI risks and how we think about it, it's consequences will still continue.

And I also want to just mention that, for example, GDPR grants and users writes to AI explanations, also there are law professors, for example, Michael Wheel from Oxford writing really great articles about why the right for explanation doesn't really grant user's rights and, quote, unquote, responsible AI.

So I think these conversations will continue to move forward and we'll see more and more things to play on. And I also want to say like some of the most outrageous AI consequences nowadays like are happening in really seemingly trivial places. Like people reading news. You're reading your friend's social media posts, right? It's not that because these inactions happen in low-risk scenarios that these AIs would not lead to big societal impacts.

MR. WEST: Yep. No, that's an important point so I appreciate your counter

perspective on that. So, Qian, there's a question about healthcare. I know you have worked a lot about this so I will direct this question to you.

The person just wants to know about the future of AI algorithms in healthcare in particular. The person seems particularly interested I gather in possible negative consequences but also possible opportunities ways that AI might help read CT scans or x-rays which they seem to be doing with greater and greater degrees of accuracy. So what is your view on AI and healthcare as we move into the future?

MS. YOUNG: That is a great question. I do think AI will improve healthcare in many ways. We are seeing not more and more AI systems are moving into practice and as people, as these systems move to practice, we are seeing more work that evaluates AI not based on the model's performance but based on doctor plus AI's performance and I think that is a really advantageous thing to do.

I do want to bring out, I think two unintended consequences that we're starting to see but I don't think are getting enough attention. One is how AI really exaggerates the existing healthcare divide between the rich and poor, between urban and rural in the U.S.

There are hospitals in rural areas. They're still not digitized. They're still using paper and pen digital records. And how can they get benefits from AI entirely. And so, I think this is the thing that larger fairness questions like how do we level the playing field is a really important question.

The second aspect I want to mention is again how do we make sure that we have a workforce, a medical workforce that can cope with this new future of AI infused healthcare? We're seeing, for example, in robotic surgery, surgeons who can operate surgical robots are really in shortage and they're really valuable. And I think that kind of trend will continue to exist.

It's not that AI spit out a prediction that means everyone can use it and when so really was used to say like AI prediction is like the P value everyone can turn out a P value using spreadsheets, but not everyone can read it correctly. And I think that's the kind of expertise we need from our future medical workforce. You need to know the meaning of an AI prediction and how to actually incorporate it in practice.

And I think that's just an example, but I think that we will be seeing large the workforce transformation because of AI in healthcare.

MR. WEST: So one closing question that I'd like each of you to answer. So we just heard about AI in healthcare. We know AI is being deployed quite substantially in finance and retain, in education and transportation and in a number of other areas.

So the question is should our solutions be sector specific or across the board? Meaning do we need regulation of AI in healthcare that are specific to healthcare? Regulations that maybe specific to the challenges of a type AI deployment in finance or retail or transportation? Or do we basically need some across-the-board regulations that would address questions, let's say, of AI bias regardless of the particular sector? How's that for a closing question?

MS. YOUNG: There couldn't have been a better one. I could start. I think I'm going to disagree with boss. I think, well, my personal way of thinking about this is I would segment responsibility according to whether it's full automation, human AI collaboration or like AI as a really supporting role as a reminder something to human.

And even within the category of human/AI collaboration, there are many different categories. Are we talking about AI working with a group of experts in healthcare? Or are we talking about AI working with an end user who may or may not care about AI or its biases, for example, in social media consumption?

I mean these are the ways I think about how to regulate energy design in AI

because the amount of human control we put on them is different. But I would be happy to hear other panelists' opinions.

MR. WEST: Okay. Meg?

MS. YOUNG: When it comes to the regulation question, I think the most important broad-based interventions we can make are to our institutional design. What are the institutions that we can build to facilitate communication between the public sector and the private sector to get under the hood of proprietary systems?

And those models will be useful across sectors. Similarly, the model vendor agreement terms, those could be widely exportable across sectors. But when it comes to the specific policies that are needed, it's going to vary sector by sector as well as by system by system.

And we need, much like Elizabeth said, ways of creating iterative flexible system design that's responsive to harms as they emerge in deployment.

MR. WEST: Okay. Thank you. Elizabeth, we'll give you the final word on this.

MS. WATKINS: Oh, my goodness. What a responsibility.

MR. WEST: Sorry for all the pressure there.

MS. WATKINS: I do find the broad risk categorization to be a very exciting step. However, I also recognize like John saying, there's a lot of questions to be made around the state of the system. Whether it's human in the loop. Whether it's an augmented decision-making system.

But I also wonder about the expertise that is already present within these sectors and they know within these regulatory sectors rather and how much the regulators know about the communities that they're serving and that they're protecting.

And so, I think the models that Meg is talking about in terms of cross-sector

transparency maybe even a new model of like cross-sector collaboration where regulators from different agencies get together and compare notes would be really exciting.

MR. WEST: Okay. Well, thank you very much. Great ideas here. I like some of the ideas in terms of public engagement, the need for agency guardrails to encourage a responsible AI developing benchmarks and metrics having impact assessments and audits to make sure they're not unanticipated consequences. And then agencies doing pilot projects where they experiment with something before they scale up and potentially affect millions of people.

So it's been a very fruitful conversation. I want to thank Meg, Qian and Elizabeth for contributing your expertise. A great job on the part of each of you. For those of you who are interested, we at Brookings write regularly about AI and emerging technologies. You can check out our writings brookings.edu.

We have a blog at tech tank where we comment on contemporary policy issues and we also we would encourage you to tune into our tech tank podcast where we discuss lots of these issues. So again, thank you to our panelists and thank you to our listening audience.

* * * * *

CERTIFICATE OF NOTARY PUBLIC

I, Carleton J. Anderson, III do hereby certify that the forgoing electronic file when originally transmitted was reduced to text at my direction; that said transcript is a true record of the proceedings therein referenced; that I am neither counsel for, related to, nor employed by any of the parties to the action in which these proceedings were taken; and, furthermore, that I am neither a relative or employee of any attorney or counsel employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

Carleton J. Anderson, III

(Signature and Seal on File)

Notary Public in and for the Commonwealth of Virginia

Commission No. 351998

Expires: November 30, 2024